Docket No. RSW920000182US1

**CLAIMS:**

What is claimed is:

1   1.   A method of selecting data sets for use with a
2   predictive algorithm, comprising:
3        generating a first distribution of a training data
4   set;
5        generating a second distribution of a testing data
6   set;
7        comparing the first distribution and the second
8   distribution to identify a discrepancy between the first
9   distribution and the second distribution; and
10        modifying selection of entries in one or more of the
11   training data set and the testing data set based on the
12   discrepancy between the first distribution and the second
13   distribution.

1   2.   The method of claim 1, wherein the first
2   distribution and the second distribution are
3   distributions of drive time from a customer geographical
4   location to a commercial establishment geographical
5   location.

1   3.   The method of claim 1, wherein the first
2   distribution and the second distribution are
3   distributions of distance between a customer geographical
4   location and a commercial establishment geograhical
5   location.

Docket No. RSW920000182US1

1   4.   The method of claim 1, wherein comparing the first
2   distribution and the second distribution includes
3   comparing one or more of a mean, mode, and standard
4   deviation of the first distribution to one or more of a
5   mean, mode, and standard deviation of the second
6   distribution.

1   5.   The method of claim 1, wherein the first
2   distribution and the second distribution are
3   distributions of a weighted distance between a customer
4   geographical location and commercial establishment
5   geographical locations.

1   6.   The method of claim 1, wherein the first
2   distribution and the second distribution are
3   distributions of a weighted drive time between a customer
4   geographical location and commercial establishment
5   geographical locations.

1   7.   The method of claim 1, wherein modifying selection
2   of entries in one or more of the training data set and
3   the testing data set includes generating recommendations
4   for improving selection of entries in one or more of the
5   training data set and the testing data set.

1   8.   The method of claim 1, wherein the training data set
2   and the testing data set are selected from a customer
3   information database.

Docket No. RSW920000182US1

1  9.   The method of claim 1, further comprising comparing
2  at least one of the first distribution and the second
3  distribution to a distribution of a customer database.

1  10.  The method of claim 1, wherein the first
2  distribution and second distribution are frequency
3  distributions of one of drive time and distance between a
4  customer geographical location and one or more commercial
5  establishment geographical locations.

1  11.  The method of claim 9, wherein comparing at least
2  one of the first distribution and the second distribution
3  to a distribution of a customer database includes:
4      generating a composite data set from the training
5  data set and the testing data set; and
6      generating a composite distribution from the
7  composite data set.

1  12.  The method of claim 1, wherein modifying selection
2  of entries in one or more of the training data set and
3  the testing data set includes changing one of a random
4  selection algorithm and a seed value for a random
5  selection algorithm.

1  13.  The method of claim 1, further comprising training a
2  predictive algorithm using at least one of the training

1 data set and the testing data set if the discrepancy is
2 within a predetermined tolerance.

1 14. The method of claim 13, wherein the predictive
2 algorithm is a discovery based data mining algorithm.

1 15. An apparatus for selecting data sets for use with a
2 predictive algorithm, comprising:
3 a statistical engine; and
4 a comparison engine coupled to the statistical
5 engine, wherein the statistical engine generates a first
6 distribution of a training data set and a second
7 distribution of a testing data set, the comparison engine
8 compares the first distribution and the second
9 distribution to identify a discrepancy between the first
10 distribution and the second distribution and modifies
11 selection of entries in one or more of the training data
12 set and the testing data set based on the discrepancy
13 between the first distribution and the second
14 distribution.

1 16. The apparatus of claim 15, wherein the first
2 distribution and the second distribution are
3 distributions of drive time from a customer geographical
4 location to a commercial establishment geographical
5 location.

1 17. The apparatus of claim 15, wherein the first
2 distribution and the second distribution are

Docket No. RSW920000182US1

1 distributions of distance between a customer geographical
2 location and a commercial establishment geograhical
3 location.

1 18. The apparatus of claim 15, wherein the comparison
2 engine compares the first distribution and the second
3 distribution by comparing one or more of a mean, mode,
4 and standard deviation of the first distribution to one
5 or more of a mean, mode, and standard deviation of the
6 second distribution.

1 19. The apparatus of claim 15, wherein the first
2 distribution and the second distribution are
3 distributions of a weighted distance between a customer
4 geographical location and commercial establishment
5 geographical locations.

1 20. The apparatus of claim 15, wherein the first
2 distribution and the second distribution are
3 distributions of a weighted drive time between a customer
4 geographical location and commercial establishment
5 geographical locations.

1 21. The apparatus of claim 15, wherein the comparison
2 engine modifies selection of entries in one or more of
3 the training data set and the testing data set by
4 generating recommendations for improving selection of
5 entries in one or more of the training data set and the
6 testing data set.

Docket No. RSW920000182US1

1   22.   The apparatus of claim 15, further comprising a
2   training data set/testing data set selection device that
3   selects the training data set and the testing data set
4   from a customer information database.

1   23.   The apparatus of claim 15, wherein the comparison
2   engine further compares at least one of the first
3   distribution and the second distribution to a
4   distribution of a customer database.

1   24.   The apparatus of claim 15, wherein the first
2   distribution and second distribution are frequency
3   distributions of one of drive time and distance between a
4   customer geographical location and one or more commercial
5   establishment geographical locations.

1   25.   The apparatus of claim 23, wherein the comparison
2   engine compares at least one of the first distribution
3   and the second distribution to a distribution of a
4   customer database by:
5       generating a composite data set from the training
6   data set and the testing data set; and
7       generating a composite distribution from the
8   composite data set.

1   26.   The apparatus of claim 15, wherein the comparison
2   engine modifies selection of entries in one or more of
3   the training data set and the testing data set by

Docket No. RSW920000182US1

4    changing one of a random selection algorithm and a seed

5    value for a random selection algorithm.


1    27.   The apparatus of claim 15, further comprising a

2    predictive algorithm device, wherein the predictive

3    algorithm device is trained using at least one of the

4    training data set and the testing data set if the

5    discrepancy is within a predetermined tolerance.


1    28.   The apparatus of claim 27, wherein the predictive

2    algorithm is a discovery based data mining algorithm.


1    29.   A computer program product in a computer readable

2    medium for selecting data sets for use with a predictive

3    algorithm, comprising:

4         first instructions for generating a first

5    distribution of a training data set;

6         second instructions for generating a second

7    distribution of a testing data set;

8         third instructions for comparing the first

9    distribution and the second distribution to identify a

10   discrepancy between the first distribution and the second

11   distribution; and

12        fourth instructions for modifying selection of

13   entries in one or more of the training data set and the

14   testing data set based on the discrepancy between the

15   first distribution and the second distribution.

Docket No. RSW920000182US1

1    30.  The computer program product of claim 29, wherein

2    the first distribution and the second distribution are

3    distributions of drive time from a customer geographical

4    location to a commercial establishment geographical

5    location.

1    31.  The computer program product of claim 29, wherein

2    the first distribution and the second distribution are

3    distributions of distance between a customer geographical

4    location and a commercial establishment geographical

5    location.

1    32.  The computer program product of claim 29, wherein

2    the third instructions for comparing the first

3    distribution and the second distribution include

4    instructions for comparing one or more of a mean, mode,

5    and standard deviation of the first distribution to one

6    or more of a mean, mode, and standard deviation of the

7    second distribution.

1    33.  The computer program product of claim 29, wherein

2    the first distribution and the second distribution are

3    distributions of a weighted distance between a customer

4    geographical location and commercial establishment

5    geographical locations.

1    34.  The computer program product of claim 29, wherein

2    the first distribution and the second distribution are

3    distributions of a weighted drive time between a customer

Docket No. RSW920000182US1

4  geographical location and commercial establishment

5  geographical locations.

1  35.   The computer program product of claim 29, wherein

2  the fourth instructions for modifying selection of

3  entries in one or more of the training data set and the

4  testing data set include instructions for generating

5  recommendations for improving selection of entries in one

6  or more of the training data set and the testing data

7  set.

1  36.   The computer program product of claim 29, further

2  comprising fifth instructions for comparing at least one

3  of the first distribution and the second distribution to

4  a distribution of a customer database.

1  37.   The computer program product of claim 29, wherein

2  the first distribution and second distribution are

3  frequency distributions of one of drive time and distance

4  between a customer geographical location and one or more

5  commercial establishment geographical locations.

1  38.   The method of claim 36, wherein the fifth

2  instructions include:

3        instructions for generating a composite data set

4  from the training data set and the testing data set; and

5        instructions for generating a composite distribution

6  from the composite data set.

Docket No. RSW920000182US1

1    39.   The computer program product of claim 29, wherein
2    the fourth instructions for modifying selection of
3    entries in one or more of the training data set and the
4    testing data set include instructions for changing one of
5    a random selection algorithm and a seed value for a
6    random selection algorithm.

1    40.   The computer program product of claim 29, further
2    comprising fifth instructions for training a predictive
3    algorithm using at least one of the training data set and
4    the testing data set if the discrepancy is within a
5    predetermined tolerance.